DOCUMENT RESUME

ED 287 507 IR 052 184

AUTHOR	Brunning, Dennis R.
TITLE	Mass Communication Thesauri and Controlled Vocabularies: An Evaluation.
PUB DATE	2 Aug 87
NOTE	15p.; Paper presented at the Annual Conference of the Association for Education in Journalism and Mass Communication (75th, San Antonio, TX, August 1-5, 1987).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Design Requirements; *Evaluation Criteria; Indexing; Journalism; *Mass Media; Professional Associations; Relevance (Information Retrieval); Subject Index Terms; *Thesauri; User Needs (Information); *Vocabulary
IDENTIFIERS	*Association for Ed in Journalism and Mass Commun

ABSTRACT

Intended as a basis for determining how the Association for Education in Journalism and Mass Communication (AEJMC) should proceed to meet its controlled vocabulary needs, this paper analyzes and evaluates the major published controlled vocabularies and thesauri used for the retrieval of mass communication literature. Following brief definitions of controlled vocabulary and thesauri, the requirements of a good controlled vocabulary are identified, as: (1) being adequate and relevant; (2) being sufficiently specific to allow a majority of searches to be conducted at an acceptable level of precision; (3) being sufficiently precoordinate to avoid false coordination and incorrect term relationships; (4) being consistent in indexing and searching through control of synonyms, near synonyms, and quasi-synonyms; (5) having reduced terminological ambiguity; and (6) helping in the selection of appropriate terms through some form of hierarchical and cross-reference structure. These six criteria are then applied to the Library of Congress Subject Headings, Communication Abstracts, Journalism Abstracts, Thesaurus: Mass Communication, and the Thesaurus of ERIC (Educational Resources Information Center) Descriptors. On the basis of this evaluation, it is concluded that AEJMC requires, at the most, its own indexing/abstracting services or, at the least, a revision of current practices, coupled with the design of a thesaurus adequate to its requirements. (Author/KM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED287507

Mass Communication Thesauri and Controlled Vocabularies: An Evaluation

Dennis R. Brunning
Communications Librarian
University Libraries
Arizona State University
Tempe, Arizona 85287

A paper presented to the Mass Communication Bibliographers Special Interest Group of the Association for Education in Journalism and Mass Communication, annual meeting, Trinity University, San Antonio, Texas, Aug. 2, 1987.

IR052164

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Dennis R. Brunning

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

This paper is about the major published controlled vocabularies and thesauri used for the retrieval of mass communication literature. I will first identity the crucial elements of a good controlled vocabulary for the indexing of our literature. I will, then, apply these criteria to assess the worth of five works: the controlled vocabularies used by the Library of Congress, Communication and Journalism Abstracts and two thesauri, the Thesaurus of ERIC Descriptors and Jean Viet's Thesaurus: Mass Communication published by UNESCO.[1] On the basis of this evaluation, I will conclude with a recommendation to this group on how we should proceed with the controlled vocabulary needs of AEJMC.

What's in a Word: The Use of Controlled Vocabularies and Thesauri[2]

Before we proceed any further, we should first become familiar and clear about what we mean by a controlled vocabulary and thesaurus. Simply put, a controlled vocabulary is an authorized and authoritative list of words used for indexing and retrieval. This list of preferred terms, as Wilfrid Lancaster describes it, controls synonyms, quasi-synonyms, and distinguishes homographs while at the same time linking together terms that are semantically related. By constructing and using a controlled vocabulary, we

avoid dispersion of like subject matter, avoid collocation of unlike subject matter and achieve a certain consistency in the representation of subject matter in indexing and searching. In other words, we just don't look anywhere using any descriptor but in certain places, with certain words--words that the structure of the controlled vocabulary itself helps us find.

A thesaurus is a specific type of controlled vocabulary. Following Lancaster again, we find that a thesaurus is essentially a limited vocabulary of terms, in alphabetical order, that can be used in indexing and searching. Like the controlled vocabulary it controls synonyms and distinguishes homographs. Unlike the controlled vocabulary, it brings its terms together in hierarchical (genus/species or broad term/narrow term) and semantic (related terms) order. As an added feature, terms whose meanings are not clear from their position in the thesaurus, are given brief definitions called scope notes.

A thesaurus, then, is a sophisticated form of the controlled vocabulary. It has all the features and functions of the controlled vocabulary but is published in a format that reveals the logical and classified order underlying a vocabulary and by extension, underlying a research field.

Numerous efforts have been made to empirically define what constitutes a good controlled vocabulary--whether in thesaurus

form or otherwise. Lancaster admits that such efforts have ranged from the significant to the futile, from the practical to the trivial. However, given the goals we have just mentioned, six important requirements seem to suggest themselves for a good controlled vocabulary:

1. It must have both literary and user warrant, that is, it must reflect the terminology used in actual texts and the terminology of the inquirer. In short it must be adequate and relevant.

2. It must be sufficiently specific to allow conduct of a great majority of searches at an acceptable level of precision. This is librarian talk which means that a term must be able to retrieve possibly relevant documents, that is, not miss the boat entirely, but it must also exclude irrelevant material, that is, be sufficiently on target.

3. It must be sufficiently precoordinate to avoid false coordination and incorrect term relationships. For example, without more information, say in the form of a subheading, the descriptor "separation anxiety" might refer to the anxiety experienced by a mother losing her child or the child losing a mother.

4. It must promote consistency in indexing and searching by control of synonyms, near synonyms, and quasi-synonyms. That is, the same concept should not be indexed by two similar descriptors or similar concepts denoted by the same descriptor.

5. It should reduce terminological ambiguity through the

separation of homographs and make clear indexing uses by scope notes.

6. Ideally, it should help indexers and searchers to select appropriate terms to represent a subject through some form of hierarchical and cross-reference structure--a thesaurus for instance.

These, then, are six criteria to consider when one evaluates a controlled vocabulary or thesaurus. Munched on whole, they are undoubtedly difficult to swallow. In my evaluations, I have kept them in mind in the following manner. In every case I have asked myself, is this list adequate? Does it reflect the textual terminology and how my colleagues and I asked questions of the literature? This questions deals with the important consideration of warrant. For the more formal questions, that is, the clarity, coherence, and cohesiveness of the vocabulary's structure, I have asked of the controlled vocabularies, to what extent can I fit this list of descriptors into a thesaurus form? For the thesauri I have asked, to what extent do these works conform to the rigorous definitions of retrieval thesauri?

Why these Works?

Before we begin evaluating our lists, I would like to thank those of you who have suggested thesauri and controlled vocabularies to evaluate. I appreciate the help and if I haven't included

everything suggested to me, I did so for reasons of time and emphasis. The controlled vocabularies under consideration here are the major published lists that ought to concern us in this, our first, foray. Should we wish to construct our own thesaurus, many more lists would be solicited and studied.

I should also note at the onset that these controlled vocabularies and thesauri were constructed for different purposes which accounts for the form they happen to take. I try to note these differences. However, for the purposes of this evaluation I have treated each list as if it were a list of preferred terms for indexing the literature written and used by AEJMC members.

I will begin with the controlled vocabularies.

Library of Congress Subject Headings

No library is without a copy of the venerable "red books" although I seriously doubt how often these works are actually "read." Nevertheless, to the extent which mass media studies are ensconced in books and to the degree to which we honor a very old and lovingly worked over list of words, we must consider the **Library of Congress Subject Headings** as a major controlled vocabulary for mass communication retrieval. Moreover, since IAC uses an enhanced version of LCSHs for its immensely popular **Infotrac**--the gateway to magazines--it behooves us to look at the adequacy of this controlled vocabulary for our needs.

The particular virtue of LCSHs is its thoroughness, both in a substantive and bibliographic sense. It is a precoordinated list of terms (that is, descriptors are linked in phrase-like strings) that attempts to record the subject matter of all writings housed at the Library of Congress, a collection to which mass communication scholars have contributed more than a few texts. Thus just about any broad topic in our field has received a descriptor or two and the richness of the standard subdivisions--descriptors for time, place, treatment--culminate in fairly accurate and relevant descriptions for retrieval. Cross references are also very instructive and helpful.

The biggest disadvantage, easily, is the general scope and lack of specificity of the headings. The Library of Congress works with a broad concept of subject; catalogers provide usually two, no more than six, descriptors for any book. In short, the headings are just not specific enough to adequately index documents, especially articles. This is particular evident when one considers IAC's use of the headings; even in enhanced form they do not nail down what an article is about and would thus be quite useless for advanced retrieval--human or machine.

Communication Abstracts

In just under ten years of its existence, **Communication Abstracts** has given us solid selective indexing of the major communication

and communication related journals and books. That it has not automated retrieval has been a major concern to many of us. However, after looking at the structure of its vocabulary and its indexing practices, I have concluded that serious work would have to be undertaken to revise the vocabulary to support online retrieval. A rich vocabulary and excellently written abstracts are assets; a very poorly constructed and maintained controlled vocabulary are definite liabilities.

By my count, in the nine years of its existence **Communication Abstracts** has used slightly over 1200 terms to index the literature. Articles are given no more than four descriptors; the descriptors are listed in a subject index with the number of the relevant citation listed. No attempt is made in this brief subject index to collocate citations. Nor is there a large effort to cross-reference terms except in the broad categories like media and television. On the whole, then, it is not much of a controlled vocabulary save for the dedicated effort to attain literary warrant.

As far as I can tell, indexers at **Communication Abstracts** work from an authority list of terms that is built from keywords in titles. These keywords are then listed in the vocabulary list and are reused when the same keyword or keyword synonym is reencountered. This is a legitimate technique, more or less, but it becomes seriously compromised when efforts are not made to

control for synonyms and other duplication of terms.

Communication Abstracts is full of such duplications. For example, one year they may use the descriptive phrase "Broadcast Satellites (Direct)" while another year they may use the phrase "Direct Broadcast Satellites." These occurrences are numerous as are the instances of using several terms to denote one concept. For example, sometimes the term "accuracy" is used to denote the concept accuracy in communication while at other times, the phrase "communication accuracy" is used. Or the notion of communicative competency is at one point referred to as "competency" while at another point the term "communication competency" is used. This inconsistency is not only logically questionable, it does not make for good and efficient retrieval if, for example, I want to track articles on competency over the years.

Communication Abstracts could also profit from some attention to hierarchical relationships. It is perhaps an artifact of terminological confusion in our field, but how many types of communication can there be? Here are some from CA: informal, formal, relational, corporate, group, small group, interpersonal, intrapersonal, marital, man-animal, office, political, public, private, personal, parent-child--the list goes on. Now I do not doubt that researchers write with such terminological panache and freedom so that we may never anticipate an end to this fissuring

of the field--I imagine irregular, impartial, indigenous and incestuous communication as probable contenders. Now perhaps it is not the task of librarians or indexers to intervene in this lexical license to say, "hey wait a minute...what are we talking about here anyway?" Such questions, thought through and properly asked, would result, I believe, in a much more ordered discourse and, subsequently, more ordered retrieval. At this point the controlled vocabulary of **Communication Abstracts** does little to help us manage these problems of disorders, the first task of any retrieval language.

Journalism Abstracts

A less ambitious yet more wisely constructed and maintained controlled vocabulary is AEJMC's list of terms used to index theses and dissertations in **Journalism Abstracts**. Beginning in 1983, under the editorship of Frances Wilhoit, **Journalism Abstracts** has maintained a published controlled vocabulary in its annual volume, listing indexing terms even if they are not used in the particular volume.

This is a very prudent list, less ambitious than **Communication Abstracts**' yet more responsible and sensitive as a controlled vocabulary. Synonyms are controlled and there is extensive use of cross references. The only drawback is the scope; limited to the indexing of theses/dissertations in journalism, it does not embrace a larger, more varied literature. At the same time,

however, it is a very solidly constructed and maintained controlled vocabulary.

Thesaurus: Mass Communication

The **Thesaurus: Mass Communication** was produced for UNESCO on the initiative of the Division of Communication Research and Planning in 1975. Compiled by Jean Viet, an important figure in French indexing and abstracting, it was originally designed for the purpose of analyzing works, articles, and other documents relating to the problems of mass communication according to the particular medium to which they refer: books, the press, records, films, radio, or television. Presently in use internally by UNESCO, parts of it can also be found in the **Unesco Thesaurus** compiled by Jean Aitchison in 1977 to serve as a retrieval vocabulary for a number of international education, science, social science, cultural, and communication agencies.

Both are classically designed thesauri. Formally, they are immaculate. They violate none of the rules. Because Aitchinson's work is more general and monolingual, it has fewer terms than Viet's; at the same time, however, its general purpose nature provides us numerous terms that often escape the strictly mass communication lists--eg. terms useful for social psychology.

Viet's vocabulary is more extensive. The vocabulary is rich enough to index, at times without nuance, most articles, say,

those indexed in **Communication Abstracts**. Coupled with its excellent thesaurus structure, it is undoubtedly a fine starting point for a more comprehensive thesaurus for mass communication retrieval.

Thesaurus of ERIC Descriptors

The **Thesaurus of ERIC Descriptors** is included for several reasons. One, in both print and online versions, ERIC is one of the more widely used and useful general indexes of social science research available. Second, as a result of excellent cooperation between ERIC and AEJMC, a number of AEJMC publications are now being indexed in the CIJE and RIE. Finally, the **Thesaurus of ERIC Descriptors** is an excellent example of thesaurus design that grapples with the problems of social science language and retrieval, a design which violates some basic rules of thesaurus construction but nevertheless does a fine job at retrieval.

Like Aitchinson's work, the **Thesaurus of Eric Descriptors** is a multipurpose social science thesaurus that includes terms relevant for mass communication retrieval. Compared to **Communication Abstracts**, it has far fewer terms, preferring more general descriptors to situate the study. For instance, both **Communication Abstracts** and the **Thesaurus of Eric Descriptors** index many of the same **JQ** articles and more often than not what ERIC retains as an identifier **Communication Abstracts** uses as a descriptor.[3] Thus using ERIC descriptors only, say in an online

search, would yield a high recall, less precise search owing to the fact that precision resides in the use of the more specific terms found in the identifier field. Since these identifier terms do not enter the official list, it limits the immediate utility of **Thesaurus of Eric Descriptors**. At the same time, however, the identifiers can be used as a rich source of additional terminology, a fortuitous advantage of the ERIC indexing system.

Recommendation

Overall, I am impressed by the foundations laid by these controlled vocabularies and thesauri. Despite **Communication Abstracts'** problems as a controlled vocabulary, it remains a rich list of terms. The **LCSHs**, though limited in scope, provides us with clues to important coordinations of subject terms with various nonsubject parameters that are useful in retrieval--intellectual treatment, physical format, etc. **Journalism Abstracts**, I will repeat, is a fine controlled vocabulary with much promise were it extended to a larger body of literature to index. Finally, the existing thesauri, while possibly not as specialized as required, are nevertheless well constructed thesauri, and are,in short, good models to follow.

Were I making recommendations to each service, I would ask it to improve its list following the criticisms I have lodged. To this group, however, I suggest that we seriously consider how we may improve upon existing vocabularies to provide ourselves with a

thesaurus that addresses our particular retrieval needs. AEJMC literature is, in some cases, covered by services using the controlled vocabularies we have critiqued; in many cases it is not covered at all. At the most, we require our own Indexing/Abstracting services; at the least, a significant revision of current practices. And in pursuit of such a service or revision, why not design a thesaurus adequate to our requirements? Knowing what we now know about existing terminological enterprises, would it not be a wise use of our talents and energies to capture the language we hear as we go about our business? Thesauri reduce verbal clutter--a task someone needs to perform. Why not us?

NOTES

1. The Library of Congress, Subject Cataloging Division produces the **Library of Congress Subject Headings, 10th edition** (Washington, DC: Library of Congress, 1986). Thomas F. Gordon edits **Communication Abstracts: An International Information Service** (Beverley Hills, CA, 1978--). The **Thesaurus of ERIC Descriptors 11th edition** (Phoenix, Oryx Press, 1987) is edited by James E. Houston. Jean Viet is the author of **Thesaurus: Mass Communication** (Paris, Unesco, 1975); the document is also available in microfiche through ERIC (ED 142 172). The two volume **UNESCO Thesaurus** (Paris, Unesco, 1977) was compiled by Jean Aitchison.

2. All comments in this section are based upon either of two works by the well known authority on controlled vocabularies, F. Wilfrid Lancaster: **Vocabulary Control for Information Retrieval,2nd edition** (Arlington, VA: Information Resources Press, 1986) and **Information Retrieval Systems: Characteristics, Testing, and Evaluation, 2nd edition** (NY: John Wiley and Sons, 1979).

3. The identifier field in ERIC includes keywords not listed in the thesaurus but are thought to be useful in online retrieval. The identifier fields, though listed in an **ERIC** citation in the print version, are not indexed.